Clinically relevant RNA-virus whole genome assembly testing on metatranscriptomic data gained directly from COVID-19 clinical samples

Dominik Hadzega, Klaudia Babisova, Patrik Krumpolec, Michaela Hyblova, Oliver Petrovic, Gabriel Minarik
MEDIREX GROUP ACADEMY, n. o., Nitra

Recent pandemics of COVID-19 showed importance of modern methods of molecular biology and bioinformatics in understanding and dealing with such events. In our project, we were sequencing nasopharyngeal swabs from COVID-19 patients. Here in this study, which is a part of mentioned bigger project, we bring brief report of testing RNA virus assembly possibilities for short read Illumina sequencing. We were dealing with metatranscriptomic data from COVID-19 positive patients. There are multiple options of assemblers for RNA genomes. We tested performance of multiple strategies of genome assembly by SPAdes and Trinity. We were evaluating and comparing quality of assemblies with help of Quast tool. We observed, that outcome of assembly is affected by chosen strategy, but we were able to construct genomes only from our short-read sequencing even *de novo* without using reference genome. Additionally, we assigned assembled genomes to SARS-CoV-2 lineage.
Keywords: RNA virus genome assembly, SPAdes, Trinity, SARS-CoV-2, RNA-seq

Skladanie genómu klinicky relevantných RNA vírusov z klinických vzoriek pochádzajúcich od COVID-19 pacientov
Počas pandémie ochorenia COVID-19 sa dostali do popredia viaceré moderné metódy molekulárnej biológie a bioinformatiky, ktoré nám pomáhali lepšie porozumieť pandémii a pomáhali s ňou bojovať. V tejto publikácii, ktorá je súčasťou väčšieho projektu sekvenovania nazoferyngálnych sterov s cieľom štúdie ochorenie COVID-19, sme sa venovali testovaniu skladania genómov RNA vírusov z krátkych čítaní z Illumina sekvenovania. Dáta od pacientov s COVID-19 analyzované v štúdii mali charakter metatranskriptómového sekvenovania. V súčasnosti je k dispozícii niekoľko softvérov vyvinutých na skladanie RNA genómov. My sme testovali úspešnosť niekoľkých stratégií skladania genómov s nástrojmi SPAdes a Trinity. Vyhodnocovali sme a porovnávali kvalitu poskladaných genómov s pomocou nástroja Quast. Ukázalo sa, že výsledok je naozaj ovplyvnený zvolenou stratégiou, pričom sme však boli z našich krátkych čítaní schopní zostaviť genómy aj na spôsob de novo bez použitia referenčného genómu. Poskladané genómy sme priradili variantom vírusu SARS-CoV-2.
Kľúčové slová: skladanie genómu RNA vírusu, SPAdes, Trinity, SARS-CoV-2, RNA sekvenovanie

Newslab, 2023; roč. 14 (1): 28 – 32

Introduction

In this study we aimed to test possibility of assembling whole genomes of RNA viruses from short reads metatranscriptomic sequencing. There are several genome assemblers that are commonly used for this purpose. Here we tested SPAdes and Trinity(1,2). However, there are other choices, for example AbySS(3), Velvet(4), IDBA(5) or Oases(6). Different assemblers vary depending on the specific characteristics of the RNA sequencing data and the species being studied. Hölzer and Marz (2019) identified several factors that affects quality of transcriptomic assembly, including read length, sequencing depth, and transcriptome complexity(7). From these results imply it is necessary to evaluate the performance of different assemblers on a case-by-case basis.

There are multiple parameters that can be evaluated quality of the assembly. One is Contiguity, which measures the length of the contigs and scaffolds produced by the assembly. Standard metrics used here are N50, L50 or N90 values. Other is completeness, measured by the proportion of the genome that has been assembled. Then there is accuracy, or in different words - how well the assembled genome matches the true sequence of the virus.

Here we assembled viral genomes directly from nasopharyngeal swabs samples of COVID-19 patients, evaluated quality of assemblies and assigned them to SARS-CoV-2 lineages.

Methods

Study Approval

Sample collection was performed as part of the clinical study approved by the Ethical committee of Bratislava Self-Governing District under the identifier 03228/2021/HF from January 12, 2021. All patients have filled the questionnaires with relevant information regarding their health status in relation to COVID-19 and signed informed consent.

Table 1. Strategies of SARS-CoV-2 assembly.

Strategy	Reads filtering	Assembler	Mode
1	mapping on SARS-CoV-2 genome	coronaspades.py	
2	not mapping on human hg38 genome	coronaspades.py	
3	not mapping on human hg38 genome	Spades	--rna
4	not mapping on human hg38 genome	Spades	--rnaviral
5	mapping on SARS-CoV-2 genome	Trinity	

Sample preparation

Samples

Nasopharyngeal swabs from COVID-19 suspected patients were gained in the two basic regimes. The patients hospitalized with middle to severe disease symptoms were recruited to the study in the cooperating hospitals. Patients with mild or any disease symptoms were recruited in the SARS-CoV-2 testing facilities of Medirex Inc. routinely collecting and analyzing samples from the population in the COVID-19 pandemic.

Nucleic acid extraction

Nasopharyngeal swabs specimens were collected from COVID-19 patients and controls and stored in viRNAtrap collection medium (GeneSpector, Czech Republic) at 4 °C. Total nucleic acid was extracted using Sera-Xtracta Virus/Pathogen Kit (Cytiva, UK) according to manufacturer instruction. 400 µl of the nasopharyngeal swab medium was used for nucleic acid extraction with final elution to 50 µl of nuclease-free water. RNA was quantified with the Qubit™ RNA High Sensitivity Assay Kit (Invitrogen). RNA isolates were stored at -80 °C.

RNA library preparation and sequencing

The metatranscriptomic libraries were prepared using KAPA RNA HyperPrep Kit with RiboErase (HMR) (Kapa Biosystems, South Africa) according to the original protocol of manufacturer. For quantity and quality control of prepared libraries a Qubit 1X dsDNA High Sensitivity Assay Kit on Qubit 3.0 (Ivitrogen) and Agilent High Sensitivity DNA Kit on Agilent 2100 Bioanalyzer (Agilent) instruments were used. Sequencing of pooled libraries was performed on NextSeq 500 and NextSeq 2000 (Illumina) platforms using 2x75 or 2x100 paired-end sequencing setup, respectively.

Data analysis

Quality Control and Data Preparation for Analysis

First step of any analysis of RNA-seq data is quality control and this step was done by FastQC v0.11.9(8). Reads were processed by Trimmomatic v0.39 (CROP:96 HEADCROP:10 LEADING:22 TRAILING:22 SLIDINGWINDOW:4:22 MINLEN:25 and our own set of adapter sequences were used in ILLUMINACLIP step)(9). Parameters were chosen according to FastQC results.

Reads mapping

Reads were mapped to the human genome hg38 by BWA-MEM algorithm of bwa v0.7.17 package(10). Reads were mapped as paired set, otherwise parameters of mapping were set to default. Same way it was done on SARS-CoV-2 genome.

Genome assembly

We performed an assembly of SARS-CoV-2 genome using coronaspades.py tool from Spades. We tested multiple strategies of assembly. In 2 strategies we proceeded with reads mapping on SARS-CoV-2 genome. Reads mapped by BWA-MEM were extracted using samtools view from samtools v1.6(11) and Picard SamToFastq from Picard v2.27.4(12) Subsequently, Spades and Trinity assemblers were applied(1,2). In other cases, we tested Spades assembler on reads that didn't map to human genome hg38. All strategies tested are shown on the **Table 1**.

For deeper evaluation and comparison of performance (resemblance to a reference genome, N50 value and other values), we used command line (Conda) instance of Quast software(13).

Scaffolds taxonomy classification

Sacaffolds were classified by Galaxy version 2.10.1. of Blast, with megablast algorithm(14). Databases NCBI NT, RefSeqRNA, SILVA rRNA and Metagenomes & metatranscriptomes DB were used as subject database. Number of hits were limited to 1.

SARS-CoV-2 lineage assignment

To assign assembled scaffolds with SARS-CoV-2 lineage, Nextstrain web-based tool was used (https://clades.nextstrain.org/, accesed on 20.04.2023). For confirmation, we used Galaxy pipeline - „Mutation calling, viral genome reconstruction and lineage/clade assignment from SARS-CoV-2 sequencing data"(15), with reads mapping on SARS-CoV-2 genom as an input. First step of the pipeline is variation analysis with key components: BWA-MEM for mapping, Lofreq for variant calling and SnpEff for variants' annotation(16,17). Then, next step is variant reporting. The third step is to generate consensus sequences and then identify SARS-CoV-2 clades/lineages by Pangolin and Nextclade(18,19).

Results

We sequenced in total 79 samples from COVID-19 positive human patients. We performed an assembly of the SARS-CoV-2 genome using Spades and Trinity assemblers. We used multiple algorithms and strategies of assembly of Spades and compared it between each other and with Trinity assembler.

First, we tested assembly with all advantages we have – using reference genome and Spades mode for assembly of SARS-CoV-2 genome. Secondly, we observed if there would be difference if we give up advantage of filtering reads by mapping on SARS-CoV-2, so all reads that didn't map on human genome were used as an input for an assembly. However, to test the ability of assembling genome of unknown RNA viruses, we needed to test another strategies - Spades modes for assembly of transcriptome or RNA viruses (spades with --rna and --rnavirus modes). We also tried another widely used transcriptome assembler Trinity. To compare its performance for our purpose to performance of Spades, we assembled genomes from reads mapping to SARS-CoV-2.

To evaluate quality of assemblies we used multiple parameters, as it is shown on the Table 2. To complexly evaluate qualities of assemblies, software Quast was used. Results from strategies using human unmapped reads instead of SARS-CoV-2 reads are influenced by RNA from other organisms – some human RNAs left in data and bacterial RNA. Because of this influence, there are additional contigs influencing quality metrics. However, "largest alignment" and "genome fraction" statistic show, that many genomes were assembled as whole or in large fragments, it is only necessary to add one more step of cleaning additional contigs. Also, this is better shown by NGA50 statistic, counting with only reference-aligned contigs. This statistic shows similar values for all strategies used. In total 42 assemblies covered at least 98% of the reference genome (some of them fragmented in more scaffolds). From most of the quality metrics it seems using mapping to the reference genome brings advantage to assembly procedure as expected. Here, some metrics were in favour of Spades, others with Trinity. Spades results were *cleaner* – with high quality and containing fewer small fragments not incorporated into genome and contigs of different origin (using coronaspades.py). It is also shown the assembly with filtering out only host reads can produce complete or almost complete genomes too. Here, it is possible to use modes for coronavirus, RNA or RNA virus. Difference is in cleanliness of the results, since non-specific assembly modes produce contigs unaligned to the reference genome. The possibility of assembly in all strategies is graphically shown in the **Figure 1**.

From scaffolds produced by SPAdes from human-unmapped reads, none of the clinically relevant viruses (other than SARS-CoV-2) were identified by Blast (scaffolds

Figure 1. Quast histograms of genome fractions showing portion of reference genome being assembled (A) and Complete genomic features in assembled genomes (B) for all strategies used for genome assembly (1 – Coronaspades on SARS-CoV-2 mapped reads, 2 – Coronaspades on human unmapped reads, 3 – Spades --rna on human unmapped reads, 4 – Spades-rnavirus on human unmapped reads, 5 – Trinity on SARS-CoV-2 mapped reads) as one of the quality metrics, shown by samples. In the graph, all at least partially assembled samples are shown. Every column in the graph represents completeness of assembly in one of the samples.

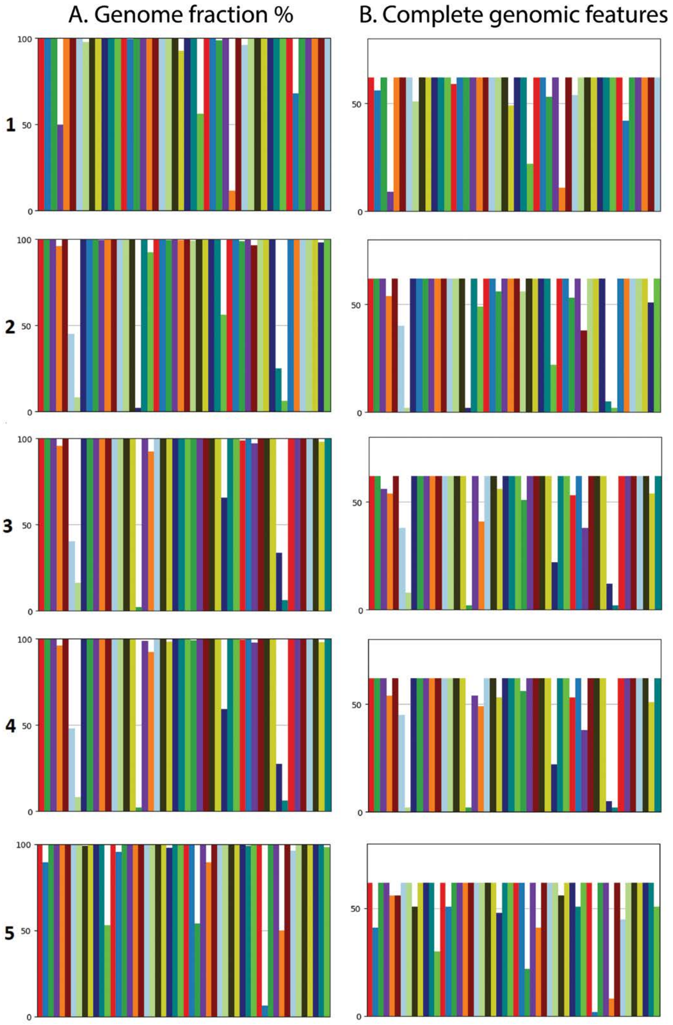
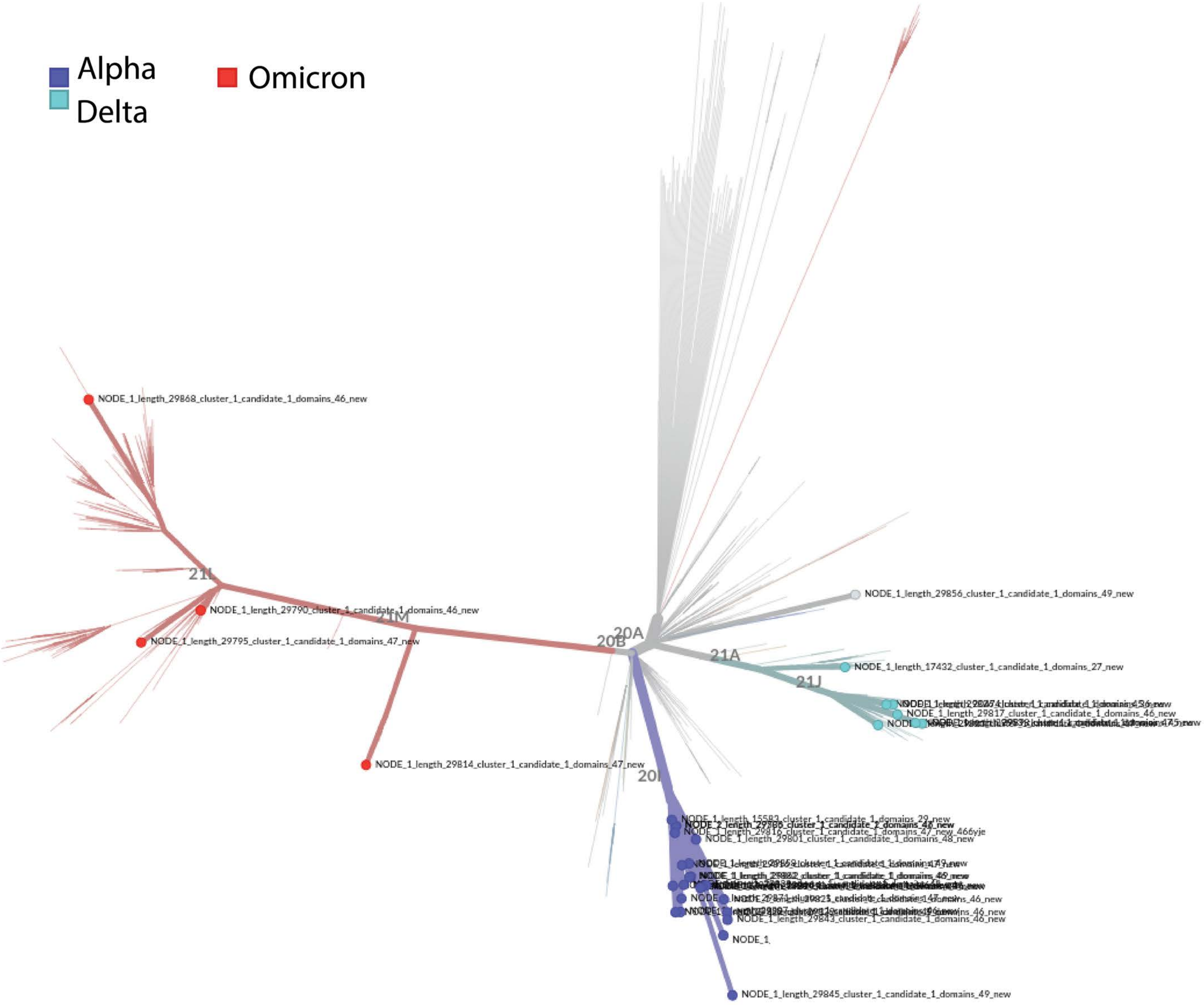


Table 2. Basic stats for assembly qualities for different strategies (As they are numbered in Table 1 in Methods section). For better understanding of statistics used by Quast, please see the authors' paper (Gurevich et al., 2013).

Mean of assembly stats	strategy1	strategy2	strategy3	strategy4	strategy5
N50	24955.39	13924.6	1438.31	1140.62	23271.15
N90	24098.5	4881.32	617.44	574.74	8583
NGA50	25968.11	26133.12	25561.3	24464.98	24282.3
NG90	26334.36	21637.39	20540.79	17102.78	25019.87
Largest alignment	25179.96	23776.4	23362.27	22285.85	23707.04
Genome fraction	94.86	89.91	90.42	90.11	94.68
GC % deviation from reference	0.16	3.05	8	7.25	0.66
Number of samples with values fitting criteria					
Genome fraction > 98	42	39	42	42	46
Genome fraction > 50	44	43	43	43	51
Largest alignment > 28 000	36	35	34	21	37
Largest alignment > 10 000	37	38	38	38	40

Figure 2. Scaffolds assignment to SARS-CoV-2 lineages / WHO variants.



🟦 Alpha
🟦 Delta
🟥 Omicron
NODE_1_length_29868_cluster_1_candidate_1_domains_46_new
21L
NODE_1_length_29790_cluster_1_candidate_1_domains_46_new
21M
NODE_1_length_29795_cluster_1_candidate_1_domains_47_new
20B
20A
NODE_1_length_29856_cluster_1_candidate_1_domains_49_new
21A
NODE_1_length_17432_cluster_1_candidate_1_domains_27_new
21J
NODE_1_length_29817_cluster_1_candidate_1_domains_46_new
NODE_1_length_29814_cluster_1_candidate_1_domains_47_new
20I
NODE_1_length_15583_cluster_1_candidate_1_domains_29_new
NODE_2_length_29786_cluster_1_candidate_2_domains_46_new
NODE_1_length_29816_cluster_1_candidate_1_domains_47_new_466yje
NODE_1_length_29801_cluster_1_candidate_1_domains_48_new
NODE_1_length_29882_cluster_1_candidate_1_domains_49_new
NODE_1_length_29843_cluster_1_candidate_1_domains_46_new
NODE_1
NODE_1_length_29845_cluster_1_candidate_1_domains_49_new

> 1000 nt were classified). Other scaffolds were mostly bacterial and human origin, some tree plants sequences aligned (Pinus strobus, Picea abies and Picea asperata). In few samples there were fragments of clinically irrelevant viruses up to 5000 nt (Pepino mosaic virus, Gallus gallus retrovirus, Apple mosaic virus, Picobirnavirus).After genome assembly we were able to look into assembled scaffolds and assign them to SARS-CoV-2 lineages. Most of the complete genomes assembled by coronaspades.py have been successfully assigned. The most frequent were variants alpha (clade 20I), then variants delta (21L) and few omicron variants (21 J). Assignment of contigs by nextclade is shown on **Figure 2.**

Discussion

This study is a part of a bigger project about human transcriptome and microbiome of COVID-19 patients. Here, on data generated for mentioned project, we tested possibility of assembling SARS-CoV-2 virus from short Illumina sequencing data with properties of metatranscriptomic sequencing (containing mainly human transcripts, but also various amounts of SARS-CoV-2 and microbial transcripts). We tested assembly strategies using advantages of available reference genome of SARS-CoV-2 or advantages of assembler mode built for the virus (coronaspades.py from SPAdes package), but also assembly ignoring these advantages to test possibilities of assembling unknown RNA virus without previously assembled genome. Results showed, that filtering reads to keep only those mapping to the virus genome helps to get *cleaner* assemblies (with higher quality, with less contigs from different organisms), so does the use of assembler mode specifically designed for SARS-CoV-2 as expected. More interesting results were about testing assembly without using reference genome to separate SARS-CoV-2 reads and with more general modes of SPAdes (--rna and –rnaviral). Here, only human reads were filtered out. Although results are messier compared to previous, it was still possible to assemble viral scaffolds covering most of the reference genome. Additional contigs were the most notable problem, although statistic NGA50 ignoring unaligned contigs showed real performance might not be so bad. Filtering final contigs from human and possibly microbial transcripts would be necessary if there wouldn't be reference for assembled virus. We also tested Trinity assembler with SARS-CoV-2 mapped reads. This assembler was not used with any mode or parameters specifically designed for SARS-CoV-2, so it had disadvantage compared to coronaspades.py, but results were still very reasonable and most parameters were comparable to SPAdes, although Quast detected few missassembled scaffolds and usually small parts of the genome were assembled fragmented.

We also validated our assemblies by subjecting them to variant assignment. We validated findings by using pipeline for galaxy environment, designed by Maier and Batut, 2023[15]. Here, we identified mix of multiple clades under WHO variant names alpha, delta, omicron and one sample with clade 20C.

From our results, we argue that de-novo assembly of RNA virus from short-read Illumina sequencing is possible, although adding long reads would be probably ideal strategy for getting more complete and reliable assemblies.

Acknowledgements

This article was created with the support of the OP Integrated Infrastructure for the project: Research on COVID-19 progressive diagnostic methods and biomarkers useful in early detection of individuals at increased risk of severe disease, ITMS: 313011ATA2, co-financed by the European Regional Development Fund.

Poďakovanie

Táto štúdia vznikala za podpory projektu: Výskum progresívnych metód diagnostiky COVID-19 a biomarkerov umožňujúcich skorú detekciu jedincov so zvýšeným rizikom ťažkého priebehu ochorenia, ITMS: 313011ATA2, spolufinancovaná Európskym fondom regionálneho rozvoja.

REFERENCES

1. Prjibelski A, Antipov D, Meleshko D, et al. Using SPAdes De Novo Assembler. Curr Protoc Bioinformatics. 2020; 70(1): e102.

2. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011; 15; 29(7): 644–652.

3. Simpson JT, Wong K, Jackman SD, et al. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009; 19(6): 1117–1123.

4. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18(5): 821–829.

5. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth, Bioinformatics. 2012; Volume 28, Issue 11, 1420–1428.

6. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012; 28(8): 1086–1092.

7. Hölzer M, Marz M. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. Gigascience. 2019; 1; 8(5): giz039.

8. Andrews S. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc; 2010.

9. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30(15): 2114–2120.

10. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv: 1303.3997v2 [q-bio.GN]; 2013.

11. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. GigaScience. 2021; Volume 10, Issue 2.

12. "Picard Toolkit." Broad Institute, GitHub Repository. Available online at: https://broadinstitute.github.io/picard/; Broad Institute. 2019.

13. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013; 15; 29(8): 1072–1075.

14. Cock PJA, Chilton JM, Grüning B, Johnson JE, Soranzo N. NCBI BLAST\mathplus integrated into Galaxy. GigaScience. 2015; 4(1).

15. Maier W, Batut B, Mutation calling, viral genome reconstruction and lineage/clade assignment from SARS-CoV-2 sequencing data (Galaxy Training Materials). Available online at: https://training.galaxyproject.org/training-material/topics/variant-analysis/tutorials/sars-cov-2-variant-discovery/tutorial. html; accessed Sun Jan 29 2023.

16. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Research. 2012; 40(22), 11189–11201.

17. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly. 2012; 6(2), 80–92.

18. Aksamentov I, Roemer C, Hodcroft EB, Neher RA. Nextclade: clade assignment, mutation calling and quality control for viral genomes. Journal of Open Source Software. 2021; 6(67), 3773.

19. O'Toole A, Scher E, Underwood A, et al. Assignment of Epidemiological Lineages in an Emerging Pandemic Using the Pangolin Tool. Virus Evolution. 2021.

Mgr. Dominik Hadžega
MEDIREX GROUP ACADEMY, *n. o.*
Novozámocká 67, 949 05 Nitra, Slovensko
e-mail: Dominik.Hadzega@medirexgroupacademy.sk